Challenges and Solutions for Latin Named Entity Recognition

Alexander Erdmann
The Ohio State University
Columbus, OH, USA
erdmann.6@osu.edu

Christopher Brown
The Ohio State University
Columbus, OH, USA
brown.2583@osu.edu

Brian Joseph
The Ohio State University
Columbus, OH, USA
joseph.1@osu.edu

Mark Janse
Ghent University
Ghent, Belgium
mark.janse@ugent.be

Petra Ajaka
The Ohio State University
Columbus, OH, USA
ajaka.3@osu.edu

Abstract

Although spanning thousands of years and genres as diverse as liturgy, historiography, lyric and other forms of prose and poetry, the body of Latin texts is still relatively sparse compared to English. Data sparsity in Latin presents a number of challenges for traditional Named Entity Recognition techniques. Solving such challenges and enabling reliable Named Entity Recognition in Latin texts can facilitate many down-stream applications, from machine translation to digital historiography, enabling Classicists, historians, and archaeologists for instance, to track the relationships of historical persons, places, and groups on a large scale. This paper presents the first annotated corpus for evaluating Named Entity Recognition in Latin, as well as a fully supervised model that achieves over 90% F-score on a held-out test set, significantly outperforming a competitive baseline. We also present a novel active learning strategy that predicts how many and which sentences need to be annotated for named entities in order to attain a specified degree of accuracy when recognizing named entities automatically in a given text. This maximizes the productivity of annotators while simultaneously controlling quality.

1 Introduction: An Overview

We present here the first evaluated Named Entity Recognition (NER) system for Latin along with the annotated data on which it was trained and tested. Our practical NER solution both caters to the unique challenges of Latin and facilitates large scale digital historiography, enabling scholars to mine the relationships of historical persons, places, and groups from a variety of primary sources. The study of historical groups specifically is a desideratum of the Herodotos project, which aims to produce a definitive catalogue of group designations in historical works (u.osu.edu/herodotos) and provided the funding for the work reported here. Using a data set drawn from the Perseus corpus (Smith et al., 2000), we develop annotation guidelines and demonstrate high inter-annotator agreement (99.3% Fleiss' Kappa). Next, we build a fully supervised NER model and evaluate it across test sets from different domains, demonstrating that it consistently outperforms baseline models regardless of how the style or register of the test set compares to the training set. Then we further address the issue of domain adaptation with an active learning solution that selects sentences to be annotated which cover gaps in the model's knowledge due to linguistic idiosyncrasies of the target domain. In addition to rapidly increasing accuracy by minimizing the amount of annotation required, a detailed error analysis demonstrates that we can reliably predict how many (and which) sentences must be annotated in a given target domain to ensure that named entities (NE) can be recognized in the remaining text with a pre-determined degree of accuracy. This provides an element of quality control for Classicists who might otherwise be wary of relying on large-scale data mining to address nuanced topics in the humanities. Finally, we discuss how the challenging qualities of Latin, being a low resource, morphologically complex language with free word order, affect our ongoing work in incorporating elements of self-training into an active learning pipeline.

2 Latin Data

The Latin language presents many challenges for NER. Being a non-standard language in terms of Natural Language Processing research, it is limited in the pre-existing training resources we can utilize. There is no annotated corpus available that makes fine-grained distinctions among NE's, and Perseus' digital gazetteers, mark-ups of Smith (1854; 1870; 1890), cover just individual persons (PRS) and geographical place names (GEO), not group names (GRP), which we want also to be able to recognize.[1] Additionally, the reliability of part-of-speech taggers (e.g. Schmid (1999) trained by Gabriele Brandolini (`http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/latin-par-linux-3.2.bin.gz`) and Johnson et al. (2014)), dependency parsers (Bamman and Crane, 2009), and semantic models (Johnson et al., 2014) is very low due to data-sparsity. Ponti and Passarotti (2016) actually demonstrate reliable syntactic dependency parsing on Medieval Latin using the *Index Thomisticus* Treebank (McGillivray et al., 2009), but find that more work is needed to successfully adapt their model to handle other varieties of Latin. In broad terms, generalizing from linguistic patterns recognized in one Latin text when processing another is difficult because "a series of particular historical, geographical and cultural circumstances [led] to an inhomogeneous linguistic system where elements from different areas and registers met and were only partially transmitted by the sources" (McGillivray, 2014). In this section, we discuss the impact of the nature of the Latin language on our selection of the Perseus corpus and our annotation of portions of it.

2.1 Digital Latin Corpora

While future work will involve incorporating diverse sources, e.g. archaeological data, liturgical Latin from *Index Thomisticus* (Busa, 1974–1980), and other historical corpora like PROIEL (Haug and Jøhndal, 2008), focusing on a single large corpus to start allows us to rely on consistent digitization. Perseus contains over 6,000,000 (mostly) consistently digitized Latin words with proper nouns (mostly) capitalized (Smith et al., 2000). This capitalization scheme did not exist in the original manuscripts but is nonetheless crucial for identifying NE's. That being said, within Perseus, the subject matter is quite diverse, certainly containing a broad spectrum of NE's of interest to Classicists. Not only does the corpus include several domains as identified by the creators – epic, elegiac, iambic and lyric poetry, tragedy, comedy, historiography, oration, mythology, philosophy, inscriptions/papyri, in addition to letters and other works (Marchionini, 2000) – but it also includes a wide breadth of diachronic variation ranging from the entire canon of Classical Latin to many early Christian authors.

2.2 Annotation

We annotated Caesar's *De Bello Gallico* (*BG*) for all the GRP's, PRS's, and GEO's appearing therein. *BG* is a fitting source text to train an NER model because it both contains many NE's and is "a model of Latin style" (Cicero, *Brutus* 262), meaning that we should be able to adapt to target texts with greater ease than if we had annotated a more peripheral, less influential source text, e.g. a work of poetry or liturgical Latin. However, because we intend to use our NER model to tag many texts which may differ considerably from *BG* in terms of content, style, and density and distribution of NE's, we chose two other texts representative of target domains widely represented in Perseus (letters and elegiac poetry) and additionally annotated portions of them as well. In total, we annotated all 58,891 words of *BG*, 18,676 of Pliny the Younger's *Epistulae* (*Ep*) and 17,562 of Ovid's *Ars Amatoria* (*AA*). *Ep* shares many more qualities with *BG* than *AA* does. Proportionally, the former share more word forms and both are prose and relatively concrete, whereas *AA* is poetic, allusive, and abstract as evidenced by the fact that a given form is far more likely to have multiple meanings in *AA* as determined by the NE labels it receives. This suggests that our target texts will allow us to define a gradient scale to measure how well we adapt to target domains based on their similarity to our source.

[1] Busa's (1974–1980) *Index Thomisticus* is annotated for proper nouns, but does not distinguish groups from places from persons as we do. Furthermore, this corpus of liturgical Latin represents a vastly different variety of the language as compared to the Classical texts we have annotated in terms of era, style, and content, so we are indeed building a novel resource. Future work will leverage *Index Thomisticus* among other resources to facilitate adapting our NER model to cover liturgical Latin.

To insure cohesive annotation among our three annotators (an undergraduate, a graduate, and a professor of Classics, each with at least 4 years of experience studying Latin), each individually annotated the same 5,000 words for NE's, over which an agreement score of 95.9% Fleiss' Kappa was achieved. Afterward, all 3 collaborated in correcting each error and the annotation guidelines were updated. Finally, another 5,000 words were individually annotated and the new agreement score of 99.3% demonstrated the consistency with which the new guidelines could be followed.

3 Fully Supervised NER Models

We built a fully supervised NER model equipped with a minimalistic, language-specific feature set and compared it to two baseline models: the only other NER technology for Latin, an unevaluated tagger from Johnson et al.'s (2014) Classical Languages Toolkit (CLTK), and a quality off-the-shelf NER model, the Stanford NER system (Finkel et al., 2005), trained on our data without language-specific features.

3.1 The Baseline Models

The CLTK NER baseline is rule based, performing a dictionary look-up and making a binary classification of entity/non-entity for each token using a gazetteer of NE's constructed by harvesting all capitalized, non-sentence-initial tokens from the Packard Humanities Institute Latin Libraries corpus (Packard Humanities Institute, 1992). Because no gold data was annotated to train or evaluate this system, our small annotated corpus presents the first opportunity to gauge its performance. Stanford's NER system is a more sophisticated conditional random field (CRF) model, although, being fully supervised in its off-the-shelf implementation, it lacks the extensive vocabulary that the CLTK system has access to. CRF's are undirected graphical models trained to maximize the conditional probability of a sequence of labels given the corresponding input sequence (Liao and Veeramachaneni, 2009), and are especially effective in NER due to their ability to rapidly learn from potentially large vectors of features belonging to each sequential token. Stanford NER leverages the widely used types of features discussed by McCallum and Li (2003), lexical, orthographic, semantic, conjoined sequences of features, and features of neighbors (we use the default feature set and parameters specified at `nlp.stanford.edu/nlp/javadoc/javanlp-3.6.0/edu/stanford/nlp/ie/NERFeatureFactory.html`), but additionally employs a Gibbs Sampling-based penalty system motivating consistency in labels for multiple occurrences of the same word type (Finkel et al., 2005).

3.2 Our Model

Our model, like Stanford's, is a CRF using similar features, though we alter ours to suit our language and corpus. We employ a POS tagger (Schmid, 1999) to leverage the highly informative morphological complexity of Latin. Finkel et al. (2005) claim only a negligible boost from using POS features in English NER and thus do not include them in the off-the-shelf version, yet we find that when implemented with creativity, the output of even a low accuracy POS tagger can be beneficial for Latin NER. For each token, we deconstruct the fine-grained POS tag into component parts ranging from case and mood distinctions to coarser distinctions between syntactic categories like nouns and verbs. We then run each token through a rule-based morphological analyzer (Whitaker, 1993), filtering out any components that the analyzer considers impossible. Furthermore, we combine the output of the tagger and analyzer with our own set of rules, thereby deducing a lemma (if one failed to be identified by the tagger or analyzer), component morphemes, and number (singular/plural), all to be used as features. Like Farber et al. (2008), which uses a similar process to leverage morpho-syntactic information in morphologically rich, low-resource Arabic, we too find that filtering POS tag output cuts down on noise, boosting accuracy. By additionally leveraging the newly enhanced LEMLAT (Budassi and Passarotti, 2016) morphological analyzer, or even substituting it for Whitaker's (1993), which has a smaller lexical base and struggles with graphical variants, we expect even further gains in coverage and accuracy with this POS tagging strategy in the future.

Further tailoring of our feature set involved tweaking parameters meant to optimize NER in data-rich environments to suit our small corpus, like limiting the size of N-grams. Lastly, we implement one fea-

Train/Test Splits

	Test Set	In or Out-of-Domain
Fold 1	PINY	OUT
Fold 2	OVID	OUT
Fold 3	CAESAR	IN

Table 1: Fold 1 tests on *Ep*, trains on the remaining annotated data. 2 tests on *AA*, and 3 on books 2 and 7 of *BG* which, when concatenated, resemble the lengths of the other 2 test sets. 3 tests "in-domain" as the training set is mostly from the same domain, i.e. the other 6 books of *BG*, the historiography domain.

Binary NE/non-NE Classification

	F	Prec	Rec	UNKF
FOLD 1 – TEST = PLINY				
CLTK	0.72	0.66	0.78	N/A
HDT-GAZ	0.96	**0.98**	0.95	0.96
HDT+GAZ	**0.97**	0.96	**0.97**	**0.96**
FOLD 2 – TEST = OVID				
CLTK	0.59	0.46	0.81	N/A
HDT-GAZ	0.89	**0.94**	0.85	0.89
HDT+GAZ	**0.91**	0.94	**0.88**	**0.91**
FOLD 3 – In-Domain TEST = GW				
CLTK	0.75	0.77	0.73	N/A
HDT-GAZ	0.99	0.99	0.98	0.97
HDT+GAZ	**0.99**	**1.00**	**0.99**	**0.98**

Full NER Classification Task

	F	Prec	Rec	UNKF
FOLD 1 – TEST = PLINY				
Stanford	0.55	0.63	0.49	0.47
HDT-GAZ	0.62	0.73	0.53	0.55
HDT+GAZ	**0.71**	**0.75**	**0.68**	**0.67**
FOLD 2 – TEST = OVID				
Stanford	0.41	0.57	0.32	0.40
HDT-GAZ	0.44	0.55	0.36	0.42
HDT+GAZ	**0.54**	**0.62**	**0.47**	**0.52**
FOLD 3 – In-Domain TEST = GW				
Stanford	0.89	0.90	0.88	0.75
HDT-GAZ	**0.91**	**0.91**	**0.91**	**0.77**
HDT+GAZ	0.91	0.91	0.91	0.76

Table 2: UNK's are words unseen in training, F, F-score, Prec, precision, Rec, recall, and GAZ, gazetteer.

ture leveraging unlabeled data which identifies any sentence-initial token which appears elsewhere in the corpus without its first letter capitalized, suggesting that it is not an NE. We find higher order semantic features like those used in Rani et al. (2014) unhelpful. We attempted to generate such features using the unevaluated Word2Vec model released with CLTK, but it could not cluster words well enough to be useful given the huge disparity between the amount of training data available in Latin and English (Mikolov et al., 2013; Michel et al., 2010).

3.3 Results

Because of the stylistic, synchronic and diachronic breadth of the Perseus corpus relative to the limited coverage of annotated data, we present both in-domain and out-of-domain results by performing 3-fold, cross-domain evaluation as described in Table 1 to assess how accuracy declines as the test domain diverges from the source. Additionally, when calculating results, we do not consider non-NE's correctly labelled in our F-score metric. This avoids artificially inflating results in low NE-dense domains (*BG* is more NE-dense than *AA* and *Ep*). Table 2 compares both the performance of CLTK NER with our model in a binary NE/non-NE classification task, and Stanford NER's performance with our model's in classifying GRP's, PRS's, and GEO's. Additionally, it displays how our model's performance is affected when aided by 3 pre-existing Latin gazetteers (Perseus' markup of Smith (1854; 1870; 1890)).

Table 2 demonstrates that the CLTK detects too many named entitites, often detecting NEs for words which can occur as both names and regular nouns or adjectives, like *Clarus*, "famous" or *Maximus*, "greatest", leading to low precision. Additionally, it cannot recognize NE's not previously stored in its dictionary, thus leading to recall errors. Both versions of our model, with and without gazetteers, outperform the Stanford baseline, proving that our language-specific POS features are helpful in coping with data sparsity. The benefit of using gazetteers, interestingly, wanes as baseline accuracy increases from Fold 2 to 1 to 3, providing no benefit in the latter. This stems from a mismatch between the gazetteers and the classification task: no gazetteer exists for GRP's and many GRP's share forms with members of the GEO gazetteer (the distinction determined by context), increasing the likelihood that these will be incorrectly tagged as GEO's. Yet as the style, topicality, and genre of the test set diverge from the training set, this problem becomes increasingly outweighed by the usefulness of having at least some gazetteer to refer to when confronted with an increased density of UNK's.

Despite our model's success in obtaining significant improvements over both baselines in all three folds, the ability to adapt to new domains is a weakness of fully supervised models. Performance drops

Accuracy vs. Density of Unknown NE's

	F	UNK's/NE's	Fold
In Domain	0.91	0.32	3
Similar Domain	0.71	0.8	1
Different Domain	0.54	0.96	2

Table 3: Decrease in accuracy as proportion of NE's in the test set not seen in training increases.

Determining the Difficulty of an UNK

	Subsets of UNK's	Accuracy
1	All UNK's (NE or non-NE)	0.96
2	Capitalized UNK's	0.77
3a	Capped UNK's appearing elsewhere uncapped	0.83
3b	Capped UNK's not appearing elsewhere uncapped	0.56

Table 4: The UNK's in row 2 are a subset of those in 1, as the UNK's in 3a and 3b are mutually exclusive, completely exhaustive subsets of those in 2. An UNK is "capitalized" if its first letter is capitalized. An UNK appears "elsewhere uncapped" if it exists somewhere in the Perseus corpus, differing only in that the first letter is not capitalized. Accuracy is reported over all 3 test sets combined.

from 91% within domain to 71% testing on a relatively similar domain (Fold 1), to 54% testing on a relatively obscure domain (Fold 2). The decline in accuracy in Table 3 suggests that the model frequently fails to leverage non-lexical features to correctly identify a label in the absence of lexical ones.

Table 4 demonstrates that forms like *Marius* (a PRS), which do not appear in the training set or possess a minimally different uncapitalized variant appearing elsewhere in Perseus, are very difficult to tag. We refer to such words as Priority 1 words. Forms like *Video*, "I see", which also do not appear in training but do show uncapitalized variants, are still challenging but much less so as most are non-NE's, only capitalized when sentence initial. We refer to these as Priority 2 words; however, these tend to be more frequent than priority 1's. We consider this tradeoff between difficulty and frequency as we tailor our pipeline to better handle capitalized UNK's.

4 Semi-Supervised Model

Semi-supervised learning involves supplementing with unannotated data during training. Liao and Veeramachaneni (2009), Rani et al. (2014), and Collins and Singer (1999) show that self-training, where unannotated data is used for training without querying the user, can overcome data sparsity or gaps between training and testing domains. However, the first two identify high precision unannotated sentences by relying on seed rules which are difficult to develop for Latin. While Liao and Veeramachaneni (2009) can rely on any capitalized word following a PRS to be part of the same NE, Latin's free word order frequently allows NE's from entirely different syntactic constituents to appear adjacent to one another, as in *Caesar [PRS] Haeduos [GRP] frumentum ... flagitare* "Caesar demands grain from the Aedui" (*BG* 1.16.1). Collins and Singer's (1999) implementation of Blum and Mitchell's (1998) co-training (the output of one tagger is used to train another) could be implemented without seed rules, yet, Pierce and Cardie's (2001) assessment of the limitations of co-training shows that when "all the classes are [not] represented according to their prior probabilities in every region in the feature space", as when we adapt to new domains in Perseus, we get Charniak's (1997) result where mistakes are magnified, not smoothed.

Active-learning, as opposed to self-training, allows the learner to query the user for additional annotation. Lynn et al. (2012) and Ambati et al. (2011) suggest that this is an effective solution for low-resource languages when self-training fails. Following Cohn et al. (1994), we modify the Query by Uncertainty tactic, where the tagger selects informative sentences based on how uncertain it is of the correct tag sequence and sends these to be annotated. Our modifications ensure that the most useful sentences are annotated first by leveraging the data-mining discussed earlier (3.3) and Sokolovska's (2011) insight that a distribution over degrees of uncertainty is more beneficial.

The sentence selection algorithm begins by identifying all Priority 1 and 2 unknown words from the test set and lists all the unannotated sentences containing at least one. Primarily, the list is ranked by

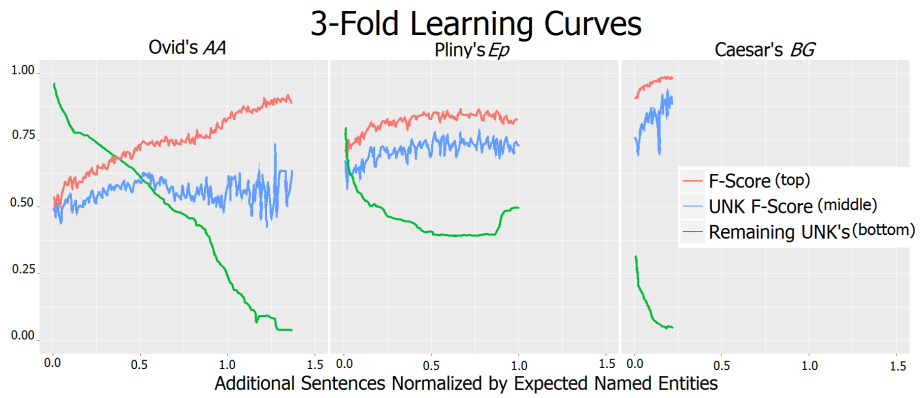

Figure 1: Learning Curves for our 3 folds demonstrate a strong negative relationship between Remaining UNK's and Accuracy. Remaining UNK's is the number of unknown words in what is left of the test set normalized by the total number of NE's therein. The increase in UNK accuracy as sentences are selected, though noisy when only few remain, demonstrates that the model learns to leverage non-lexical features as data is added to the training set.

frequency of the relevant UNK weighted by priority level, reflecting that we are more likely to mistag Priority 1's than Priority 2's and that fixing high frequency mistakes is more important than low-frequency ones (sentences with multiple unknown words can be entered into the list multiple times but only extracted once). Secondarily, we rank sentences in the list by the sum of (a) the marginal probability with which the fully supervised tagger predicts the relevant UNK's tag, and (b) the median of these marginal probabilities over every occurence of that word form in the test set. A low sum of these two outranks higher sums, *ceteris paribus*, as this implies that the given sentence is highly informative and that the UNK it features is not likely to be tagged correctly a priori. The selection algorithm then progresses through this ranked list taking sentences out for annotation only if the relevant UNK therein has not already been added to the training set by a higher-ranking sentence. In order to ensure that, during training, the tagger optimally learns from sentences which were selected by the algorithm, we weight these sentences in the training data according to the fraction of all Priority 1 and 2 types from the test set which are represented therein. Intuitively, the algorithm addresses the sources of errors discussed previously, though one flaw (4.2) is still being addressed.

4.1 Experiment

We set out to determine if (a) our sentence selection algorithm is efficient and (b) we can reliably predict tagging accuracy based on how many sentences we have already selected and annotated. Such are the practical concerns of e.g. Classicists studying the portrayal of GRPs in the liturgical *Index Thomisticus* (Busa, 1974–1980). Accuracy on a held-out set does not concern them, only how many and which sentences must be annotated within *Index Thomisticus* to ensure that the remainder can be tagged with sufficient accuracy to meet their projects' needs. Thus, we return to our 3 domain-disjoint folds over which we tested the fully supervised model, pretending that the test sets represent never-before-seen documents of varying tagging difficulties. For each fold, we run our sentence selection algorithm on the test set, incrementally updating the training set with selected sentences and testing on those remaining. Figure 1 depicts the results from running the tagger on each fold until the learning curve levels off.

The accuracy on *BG* levels off at the top of the probability space once all capitalized UNK's have been seen. *AA* leveling off in the low 90's is an effect of the inherent challenge of tagging a text in which the same NE is frequently used to refer to different classes, merely reflecting the lower inter-annotator agreement in this test set. The flaw with the sentence selection algorithm is that Remaining UNK's never approach 0 in *Ep*. The algorithm fails to select sentences for some unknown PRS's because several (*Maximus*, *Clarus*, etc.) are homonymous with non-NE's, representing a third priority to be incorporated in the updated algorithm. However, for now, we can ignore the effect of homonymy by only considering results in *Ep* from 0 to 0.5 additional normalized sentences, the other texts being free of such effects.

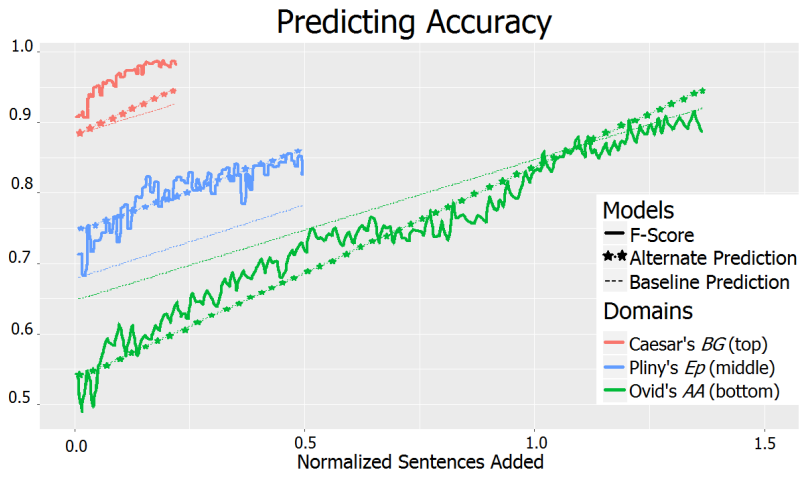

Figure 2: Leveraging distributional statistics of UNK's distinguishes the alternative model.

4.2 Predicting Accuracy *Ex Ante*

We can use linear regression[2] to predict how accurately a text can be tagged based on how many sentences have been selected, enabling our Classicists to deduce how many sentences they must annotate to ensure sufficient tagging accuracy to support their work. We compare two potential regression models:

$$Y_b = \beta_0 + \beta_1 X_1 + \epsilon \tag{1}$$

$$Y_a = \beta_0' + \beta_1' X_1 + \epsilon' \tag{2}$$

The baseline Equation 1 predicts accuracy such that β_0, the accuracy before sentences are selected, is a function of the number of capitalized UNK's normalized by the expected number of NE's in the test set.[3] The rate at which accuracy changes, β_1, as sentences (normalized by the number of expected NE's) are added, X_1, is assumed to be constant across all train/test splits. The alternative Equation 2 predicts β_0' to be a function of the density of capitalized UNK's *weighted* by the proportions of priority levels represented therein, just as the sentence selection algorithm weights priority 1 words over priority 2's. β_1' is uniquely determined for every test set based on the type-to-token ratio of capitalized UNK's (also weighted by priorities), such that a higher type-to-token ratio predicts greater accuracy gains per sentence annotated. Figure 2 demonstrates that priority levels and type-to-token frequency ratios enhance our ability to predict accuracy gains via active learning and that accuracy can be well modeled through the upper-middle regions of the probability space. Ongoing work is addressing the effect of homonymy as well as the effect of similarity in distributions over NE labels between training and test sets – this is contributing to the greater-than-predicted accuracy in *BG*.

5 Conclusion

We present an active-learning NER pipeline for low-resource Latin that expedites accuracy gains at low annotation cost. Our pipeline enables researchers to gauge the reliability of NER output and upgrade that reliability until it meets their standards for a given application. While still in development, our product makes novel contributions to the field including the first annotated corpus for evaluating NER in Latin and a formula for predicting tagging accuracy throughout the active learning process. In the coming months, we will make all of this publicly available: our annotated corpus and the supervised NER model

[2] As we are in probability space, logistic regression would seem more fitting, but there are many complex non-linear effects due to e.g. the improvement of UNK accuracy as the density of UNK's decreases, such that we are still working on developing an adequately complex regression model; however, linear regression provides very suitable predictions of accuracy as texts improve through the upper-middle ranges of the probability space with which we are chiefly concerned.

[3] The number of NE's expected in the test set is the number of non-sentence-initial capitalized words plus the number of sentences which should start with an NE given equal distributions of NE's within sentences; however, this assumption is not entirely valid and we slightly over-predict NE's, though the minor effect should not bias our results.

trained on it as well as an interface for guiding annotators through the active learning strategy described above.

Acknowledgements

The authors of this paper are indebted to Micha Elsner and Marie-Catherine de Marneffe as well as two anonymous reviewers for their invaluable input and advice.

References

Vamshi Ambati, Stephan Vogel, and Jaime G. Carbonell. 2011. Multi-Strategy Approaches to Active Learning for Statistical Machine Translation. *Proceedings of the 13th Machine Translation Summit*.

David Bamman and Gregory Crane. 2009. Structured Knowledge for Low-Resource Languages: The Latin and Ancient Greek Dependency Treebanks. Tufts University.

Avrim Blum and Tom Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-Training. *COLT 98: Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 92–100.

Marco Budassi and Marco Passarotti. 2016. *Nomen Omen*: Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon. *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, Berlin, Germany, 2016, pages 90–94.

Roberto Busa. 1974–1980. *Index Thomisticus*. Stuttgart-Bad Canstatt: Frommann-Holzboog. `www.corpusthomisticum.org/it/index.age`.

Eugene Charniak. 1997. Statistical Parsing with a Context-Free Grammar and Word Statistics. *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 598–603. Cambridge, MA; MIT Press.

Dag T. T. Haug and Marius L. Jøhndal. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In Caroline Sporleder and Kiril Ribarov (eds.), *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34.

David Cohn, Les Atlas, and Richard Ladner. 1994. Improving Generalization with Active Learning. *Machine Learning*, 15.2: 201–222.

Michael Collins and Yoram Singer. 1999. Unsupervised Models for Named Entity Classification. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Jenny R. Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370.

Benjamin Farber, Dayne Freitag, Nizar Habash, and Owen Ranbow. 2008. Improving NER in Arabic Using a Morphological Tagger. In: N. Calzolari et al. (eds.), *Proceedings of the Language Resources and Evaluation Conference (LREC'08)*, pages 2509–2514.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33: 613–619.

Kyle P. Johnson et al. 2014-2016. CLTK: The Classical Language Toolkit. DOI 10.5281/zenodo.60021.

Wenhui Liao and Sriharsha Veeramachaneni. 2009. A Simple Semi-supervised Algorithm for Named Entity Recognition. *Proceedings of the NAACL HLT Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 58–65. Boulder, CO: Association for Computational Linguistics.

Teresa Lynn, Jennifer Foster, Mark Dras, and Elaine Ul'Dhonnchandha. 2012. Active Learning and the Irish Treebank. *Proceedings of ALTA*.

Gary Marchionini. 2000. Evaluating Digital Libraries: A Longitudinal and Multifaceted View. University of North Carolina at Chapel Hill.

Andrew McCallum and Wei Li. 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. *Proceedings of the Seventh Conference on Natural Language Learning (CONLL) at HLT-NAACL*, pages 188–191. NAACL Press.

Barbara McGillivray. 2014. *Methods in Computational Linguistics*. Leiden: E.J. Brill.

Barbara McGillivray, Marco Passarotti and Paolo Ruffolo. 2009. The *Index Thomisticus* Treebank Project: Annotation, Parsing and Valency Lexicon. *TAL*, 50.2: 103–127.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2010. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 311.6014: 176–182.

Thomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*.

Packard Humanities Institute. 1992. *PHI CD ROM Format Description*. Los Altos, CA.

David Pierce and Claire Cardie. 2001. Limitations of Co-Training for Natural Language Learning from Large Datasets *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Edoardo Martia Ponti and Marco Passarotti. 2016. *Differentia Compositionem Facit*: A Slower-Paced and Reliable Parser for Latin. *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 683–688.

Pratibha Rani, Vikram Pudi, and Dipti Sharma Misra. 2014. TagMiner: A Semisupervised Associative POS Tagger Effective for Resource Poor Languages. In: P. Cellier et al. (eds.): *Proceedings of DMNLP, Workshop at ECML/PKDD*, Nancy, France, 2014, pages 113–128.

Helmut Schmid. 1999. Improvements in Part-of-Speech Tagging with an Application to German. In: Susan Armstrong et al. (eds.): *Text, Speech and Language Technology Natural Language Processing Using Very Large Corpora*, pages 13–25. Dordrecht: Kluwer Academic Publishers.

David A. Smith, Jeffrey A. Rydberg-Cox, and Gregory R. Crane. 2000. The Perseus Project: a Digital Library for the Humanities. *Literary and Linguistic Computing*, 15: 15–25.

William Smith. 1854. *Dictionary of Greek and Roman Geography*. Perseus Project.

William Smith. 1870. *Dictionary of Greek and Roman Antiquities*. Perseus Project.

William Smith. 1890. *Dictionary of Greek and Roman Biography and Mythology*. Perseus Project.

Nataliya Sokolovska. 2011. Aspects of Semi-Supervised and Active Learning in Conditional Random Fields. *Proceedings of the European Conference on Machine Learning (ECML PKDD)*, pages 273–288. Berlin: Springer.

William Whitaker. 1993. William Whitaker's Words: `http://archives.nd.edu/words.html`.